International Journal of Approximate Reasoning 50 (2009) 867–879

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Measures of ruleset quality for general rules extraction methods ☆

Martin Holeňa

Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 18207 Praha 8, Czech Republic

ARTICLE INFO

Article history:
Received 15 April 2008
Received in revised form 18 February 2009
Accepted 4 March 2009
Available online 13 March 2009

Keywords:
Rules extraction from data
Quality measures
Ruleset measures
ROC curves
Observational logic
Fuzzy logic

ABSTRACT

The paper deals with quality measures of whole sets of rules extracted from data, as a counterpart to more commonly used measures of individual rules. It sketches the typology of rules extraction methods and of their rulesets, and recalls that quality measures for whole sets of rules have been so far used only in the case of classification rulesets. Then three particular approaches to extending ruleset quality measures from classification to general rulesets are discussed. The paper also recalls the possibility to measure the dependence of classification rulesets on parameters of the classification method by means of ROC curves, and proposes a generalization of ROC curves to general rulesets. Finally, the approach is illustrated on rulesets extracted with four important rules extraction methods from the well-known iris data.

1. Introduction

Logical formulas of specific kinds, usually called *rules*, are a traditional way of formally representing knowledge. Therefore, it is not surprising that they are also the most frequent representation of the knowledge discovered in data mining. Existing methods for rules extraction are based on a broad variety of paradigms and theoretical principles. However, methods relying on different underlying assumptions can lead to the extraction of different or even contradictory rulesets from the same data. Moreover, the set of rules extracted with a particular method can substantially depend on some tunable parameter or parameters of the method, such as significance level, thresholds, dimensions, trade-off coefficients, etc. For that reason, it is desirable to have measures of various qualitative aspects of the extracted rulesets. So far, such measures are available only for sets of classification rules, and their dependence on tunable parameters can be described only for classification into two classes [1,2]. As far as more general kinds of rules are concerned, measures of quality have been proposed only for individual rules [3–9], or for contrast sets of rules, which finally can be replaced with a single rule [10,11]; if a whole ruleset is taken into consideration, then only as a context for measuring the quality of an individual rule [12,13].

This paper, which is an extended version of a talk at the ECSQARU 2007 conference [14], discusses three possible ways of generalizing existing ruleset quality measures from classification to general rulesets, as well as a generalization of the ROC curves, which have been used for studying the dependence of classification rulesets on method parameters. The proposed generalizations are introduced in Section 5, after some preliminaries from fuzzy logic in the narrow sense are recalled in Section 2, the typology and examples of rules extraction methods in Section 3, and examples of measures for classification rulesets in Section 4. The paper concludes with an illustration of the proposed approaches on the well-known iris data.

☆ The research reported in this paper has been supported by the Grant No. 201/08/0802 of the Grant Agency of the Czech Republic, and partially supported by the Institutional Research Plan AV0Z10300504. Highly appreciated is the contribution of Vojta Hlaveš, who has run all the tests.

E-mail address: martin@cs.cas.cz

0888-613X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2009.03.002

2. Fuzzy logic preliminaries

In this section, some technical details pertaining to fuzzy logic are introduced, which will be needed in various parts of the paper. This allows not to be burdened with them in the sequel, and to concentrate on ruleset quality measures instead. At the same time, gathering all relevant fuzzy logic concepts at one place reveals more clearly that fuzzy logic is used in the narrow sense in this paper [15], not in the broad sense of the fuzzy set theory.

Recall [16] that formulas of any fuzzy logic are built from *variables* (propositional or object variables), the *truth constant* $\bar{0}$ and the *connectives conjunction* $\&$ and *implication* $\rightarrow$, possibly also *quantifiers* $\forall$ and $\exists$ (if a fuzzy predicate logic is considered), and the *truth evaluation* $\|\varphi\|$ of a formula φ is a value from $[0,1]$, determined according to the following rules:

(i) the truth evaluation $\|\varphi\|$ depends on the values of the variables involved in the formula φ (except for variables bound by quantifiers); if a particular variable x is considered, then this dependence is frequently reflected using the notation $\|\varphi\|_x$;
(ii) $\|\bar{0}\| = 0$;
(iii) $\|\varphi_1 \& \varphi_2\| = \|\varphi_1\| t \|\varphi_2\|$, where t is a continuous t-norm; different t-norms determine different kinds of conjunctions;
(iv) $\|\varphi_1 \rightarrow \varphi_2\| = \|\varphi_1\| \Rightarrow_t \|\varphi_2\|$, where $\Rightarrow_t$ is the residuum of a continuous t-norm t, i.e., $x \Rightarrow_t y = \max\{z \in [0,1] : xtz \leqslant y\}$; different t-norms determine different implications;
(v) if a predicate logic is considered and a variable x has not been bound in φ by a quantifier, then $\|\forall x \varphi\| = \inf \|\varphi\|_x$ and $\|\exists x \varphi\| = \sup \|\varphi\|_x$, where the infimum and supremum are taken over all possible values of x.

Besides conjunctions and implications, formulas of a fuzzy logic frequently contain various *derived connectives*, which are used to shorten other, more complex formulas. Their best known examples are the *negation* $\neg$, with $\neg\varphi$ being a shorthand for $\varphi \rightarrow \bar{0}$, and *equivalence* $\equiv$, with $\varphi_1 \equiv \varphi_2$ being a shorthand for $(\varphi_1 \rightarrow \varphi_2)\&(\varphi_2 \rightarrow \varphi_1)$. Similarly, an additional *truth constant* $\bar{1}$ is used to shorten the formula $\bar{0} \rightarrow \bar{0}$, for which $\|\bar{0} \rightarrow \bar{0}\| = 1$, due to (ii) and (iv).

If all conjunctions and implications in all formulas of a fuzzy logic are of the same kind determined by a t-norm t, then due to (iii)–(iv) this t-norm determines the truth evaluation $\|\varphi\|$ of every formula of the logic which can be reflected in using the notation $\|\varphi\|_t$ for that evaluation. In the sequel, only such fuzzy logics will be considered. In particular, they include the three fundamental fuzzy logics [16]:

- the *Gödel logic*, with $t = \wedge$, the usual minimum of reals;
- the *Łukasiewicz logic*, with $t = *_Ł$, the *Łukasiewicz t-norm*, which is defined $x *_Ł y = \max(0, x + y - 1)$;
- the *product logic*, with $t = *$, the usual product of reals.

Among the plethora of results concerning formulas of fuzzy logic and their truth evaluations, several will be needed to prove Proposition 1 in Section 6. They are gathered in the following two lemmas, concluding this section.

Lemma 1. *Let φ be an arbitrary formula of a fuzzy logic, and let all conjunctions $\&$ and implications $\rightarrow$ in formulas of the logic be determined by a t-norm t. Recalling that a number $x \in [0,1]$ is called idempotent element of t if $xtx = x$, denote*

$$\mathscr{I}_t = \{[a,b] \subset [0,1] : b > a \& a, b \text{ are idempotent elements of } t \\ \& \ (\forall x \in (a,b)) x \text{ is not an idempotent element of } t\}. \tag{1}$$

Then

(a) *$\varphi\&\neg\varphi \rightarrow \bar{0}$, which in terms of truth evaluations means $\|\varphi\|_t \|\neg\varphi\|_t = 0$;*
(b) *$(\forall x, y \in [0,1])$ if $(\forall I \in \mathscr{I}_t)\{x, y\} \not\subset I$, then $xty = x \wedge y$;*
(c) *the restriction (I,t) of $([0,1],t)$ to an interval $I \in \mathscr{I}_t$ is isomorphic either to $([0,1],*)$ or to $([0,1],*_Ł)$;*
(d) *in the Łukasiewicz logic, $\neg\neg\varphi \equiv \varphi$, which in terms of truth evaluations means $\|\neg\neg\varphi\|_{*_Ł} = \|\varphi\|_{*_Ł}$;*
(e) *in Gödel, Łukasiewicz and product logic, the negation is evaluated*

$$\|\neg\varphi\|_{*_Ł} = 1 - \|\varphi\|_{*_Ł}, \|\neg\varphi\|_{\wedge} = \|\neg\varphi\|_{*} = \begin{cases} 1 & \text{if } \|\varphi\| = 0, \\ 0 & \text{else}. \end{cases} \tag{2}$$

Proof. (a) The tautology $\varphi\&\neg\varphi \rightarrow \bar{0}$ has been proved in Lemma 2.2.12. of [16]. Since it is a tautology,

$$1 = \|\varphi\&\neg\varphi \rightarrow \bar{0}\| = \max\{z \in [0,1] : \|\varphi\| t \|\neg\varphi\| tz \leqslant 0\}, \tag{3}$$

which implies $\|\varphi\| t \|\neg\varphi\| = 0$.
(b) and (c) have been proved in Theorem 2.1.16 of [16].
(d) The tautology $\neg\neg\varphi \equiv \varphi$ has been proved in Lemma 3.1.1 of [16]. An equivalence is a tautology if and only if the truth evaluations of both equivalent formulas are the same, hence $\|\neg\neg\varphi\|_{*_Ł} = \|\varphi\|_{*_Ł}$.
(e) Computing the residuum of $\wedge$, $*_Ł$ and $*$ yields:

$$x \Rightarrow_{\wedge} y = \begin{cases} 1 & \text{if } x \leqslant y, \\ y & \text{if } x > y, \end{cases} \tag{4}$$

$$x \Rightarrow_{*_{\text{Ł}}} y = \begin{cases} 1 & \text{if } x \leqslant y, \\ 1 - x + y & \text{if } x > y, \end{cases} \tag{5}$$

$$x \Rightarrow_{*} y = \begin{cases} 1 & \text{if } x \leqslant y, \\ \frac{y}{x} & \text{if } x > y. \end{cases} \tag{6}$$

Applying (4)–(6) to $x = \|\varphi\|$, $y = 0$ yields (2). □

Lemma 2. *Let φ, t and $\mathscr{I}_t$ have the same meaning as in Lemma 1. Then there exist a constant $c \in (0,1]$ and a strictly increasing bijective mapping $g : [0,c] \to [0,1]$ such that*

$$\|\varphi\|_t + \|\neg\varphi\|_t \leqslant \max\left(\max_{x\in[0,c]}(x + g^{-1}(1 - g(x))), 1\right). \tag{7}$$

Proof. If $\|\varphi\|_t$ and $\|\neg\varphi\|_t$ do not belong to the same $I \in \mathscr{I}_t$, then according to Lemma 1(a) and (b), $\|\varphi\|_t \wedge \|\neg\varphi\|_t = \|\varphi\|_t t \|\neg\varphi\|_t = 0$, which entails

$$\|\varphi\|_t + \|\neg\varphi\|_t \leqslant 1. \tag{8}$$

On the other hand, if $\|\varphi\|_t, \|\neg\varphi\|_t \in I \in \mathscr{I}_t$, then the fact that $\min(I)$ is an idempotent element of t implies

$$0 \leqslant \min(I) = \min(I) t \min(I) \leqslant \|\varphi\|_t t \|\neg\varphi\|_t = 0, \tag{9}$$

thus $\min(I) = 0$, i.e., $I = [0,c]$ with $c > 0$. First, consider the case that there exists an isomorphism i_P of (I,t) onto $([0,1],*)$. Then

$$\|\varphi\|_* * \|\neg\varphi\|_* = i_P(\|\varphi\|_t) * i_P(\|\neg\varphi\|_t) = i_P(\|\varphi\|_t t \|\neg\varphi\|_t) = i_P(0) = 0. \tag{10}$$

Hence, $\min(\|\varphi\|_*, \|\neg\varphi\|_*) = 0$, which implies also $\min(\|\varphi\|_t, \|\neg\varphi\|_t) = 0$, entailing again the inequality (8). Finally, if there exists an isomorphism $i_{\text{Ł}}$ of (I,t) to $([0,1], *_{\text{Ł}})$, then due to Lemma 1(d),

$$i_{\text{Ł}}(\|\neg\neg\varphi\|_t) = \|\neg\neg\varphi\|_{*_{\text{Ł}}} = \|\varphi\|_{*_{\text{Ł}}} = i_{\text{Ł}}(\|\varphi\|_t), \quad \text{thus} \|\neg\neg\varphi\|_t = \|\varphi\|_t. \tag{11}$$

Since φ was an arbitrary formula fulfilling $\|\varphi\|_t, \|\neg\varphi\|_t \in [0,c]$, (11) implies that the truth evaluation of negation corresponding to the restriction of t to $[0,c]$ is an involution. According to Proposition 2.38 of [17], there exists a strictly increasing bijective mapping $g : [0,c] \to [0,1]$ such that for any φ with $\|\varphi\|_t, \|\neg\varphi\|_t \in [0,c]$,

$$\|\neg\varphi\|_t = g^{-1}(1 - g(\|\varphi\|_t)) \leqslant \max_{x\in[0,c]}(g^{-1}(1 - g(x))). \tag{12}$$

Combining (8) and (12) already yields (7). □

3. Methods for the extraction of rules from data

3.1. Typology of rules extraction methods

The most natural base for differentiating between existing rules extraction methods is the *syntax and semantics of the extracted rules*. Syntactical differences between them are, however, not very deep since principally, any rule r from a ruleset $\mathscr{R}$ has one of the forms $S_r \sim S'_r$, or $A_r \to C_r$, where S_r, S'_r, A_r and C_r are formulas of the considered logic, and $\sim$, $\to$ are symbols of the language of that logic. The difference between both forms concerns semantic properties of the symbols $\sim$ and $\to$: $S_r \sim S'_r$ is symmetric with respect to S_r, S'_r in the sense that its validity always coincides with that of $S'_r \sim S_r$ whereas $A_r \to C_r$ is not symmetric with respect to A_r, C_r in that sense. In the case of a propositional logic, $\sim$ and $\to$ are the connectives equivalence ($\equiv$) and implication, respectively, whereas in the case of a predicate logic, they are generalized quantifiers. To distinguish the formulas involved in the asymmetric case, A_r is called *antecedent* and C_r *consequent* of r.

More important is the semantic of the rules (cf. [6]), especially the difference between *rules of the Boolean logic* and *rules of a fuzzy logic*. Due to the semantics of Boolean and fuzzy formulas, the former are valid for crisp sets of objects, whereas the validity of the latter is a fuzzy set on the universe of all considered objects. Boolean rulesets are extracted more frequently, especially some specific types of them, such as *classification rulesets* [1,8]. Those are sets of implications such that $\{A_r\}_{r\in\mathscr{R}}$ and $\{C_r\}_{r\in\mathscr{R}}$ partition the set $\mathscr{O}$ of considered objects, where $\{\cdot\}_{r\in\mathscr{R}}$ stands for the set of distinct formulas in $(\cdot)_{r\in\mathscr{R}}$. Abandoning the requirement that $\{A_r\}_{r\in\mathscr{R}}$ partitions $\mathscr{O}$ (at least in the sense of a crisp partitioning) allows to generalize those rulesets also to fuzzy antecedents [18]. For Boolean antecedents, however, this requirement entails a natural definition of the validity of a whole classification ruleset $\mathscr{R}$ for an object x. Assuming that all information about x conveyed by $\mathscr{R}$ is conveyed by the single rule r covering x (i.e., with A_r valid for x), the validity of $\mathscr{R}$ for x can be defined to coincide with the validity of $A_r \to C_r$ for that

r, which in turn equals the validity of C_r for x. Needless to say, that definition is not applicable if there exists an object covered by two rules with different consequents, even not in the case when the consequents are assignments to classes.

As far as Boolean predicate logic is concerned, generalized quantifiers both for symmetric and for asymmetric rules were studied in the 1970s within the framework of the *observational logic* [19], which is a Boolean predicate logic with generalized quantifiers. For a set of data about n objects, the truth evaluation of the Boolean predicate φ on those objects is a vector $\|\varphi\| \in \{0,1\}^n$, whereas the truth evaluation of a formula $(Qx)(\varphi_1(x),\ldots,\varphi_m(x))$ consisting of m Boolean predicates $\varphi_1,\ldots,\varphi_m$ and an m-ary generalized quantifier Q is the function value

$$\|(Qx)(\varphi_1(x),\ldots,\varphi_m(x))\| = \mathrm{Tf}_Q(\|\varphi_1\|,\ldots,\|\varphi_m\|) \tag{13}$$

of a $\{0,1\}$-valued function Tf_Q on the set of m-column binary matrices, which is called *truth function* of the quantifier Q. Observational logic underlies one of the earliest methods for the extraction of general rules from data, called General Unary Hypotheses Automaton (GUHA). In GUHA, the truth function Tf_Q of a generalized quantifier Q is always a function of the 4-fold table

		S'_r	$\neg S'_r$
		C_r	$\neg C_r$
S_r	A_r	a	b
$\neg S_r$	$\neg A_r$	c	d

. (14)

Hence, Tf_Q is a $\{0,1\}$-valued function on quadruples of non-negative integers. For symmetric rules, GUHA uses quantifiers fulfilling

$$a' \geqslant a \,\&\, b' \leqslant b \,\&\, c' \leqslant c \,\&\, d' \geqslant d \,\&\, \mathrm{Tf}_Q(a,b,c,d) = 1 \rightarrow \mathrm{Tf}_Q(a',b',c',d') = 1. \tag{15}$$

They are called *associational quantifiers*. For asymmetric rules, it uses quantifiers fulfilling the stronger condition

$$a' \geqslant a \,\&\, b' \leqslant b \,\&\, \mathrm{Tf}_Q(a,b,c,d) = 1 \rightarrow \mathrm{Tf}_Q(a',b',c',d') = 1, \tag{16}$$

and are called *implicational quantifiers*. This condition covers also the frequently encountered *association rules* [6,20–22] (since methods for the extraction of association rules have been developed outside the framework of observational logic, the terminology is a bit confusing here: although associational rules are asymmetric, their name evokes the quantifiers for the symmetric ones).

Orthogonally to the typology according to the semantics of the extracted rules, all extraction methods can be divided into two large groups:

- Methods that extract logical rules from data *directly*, without any intermediate formal representation of the discovered knowledge. Such methods have always formed the mainstream of the extraction of Boolean rules: from the observational logic methods [19] and the method *AQ* [23,24] in the late 1970s, through the extraction of association rules [20–22] and the method *CN2* [25], relying on a paradigm similar to that of AQ, to methods based on *rough sets* [26,27], *inductive logic programming* [28,29] and *genetic algorithms* [30]. They include also important methods for fuzzy rules, in particular *ANFIS* [31,32] and *NEFCLASS* [33,34], *fuzzy generalizations* of *observational logic* [35,36] and a recent method based on the theory of *evaluative linguistic expressions* [37].
- Methods that employ some *intermediate representation* of the extracted knowledge, useful by itself. This group includes two important kinds of methods: *classification trees* [38,39] and methods based on *artificial neural networks (ANN)*. The latter are used both for Boolean and for fuzzy rules [40–42] (cf. also the survey papers [43,44]).

3.2. Important examples of rules extraction methods

In this subsection, the basic principles of four important rules extraction methods will be recalled. Their choice attempts to reflect the various aspects of the differences within the spectrum of the existing methods. In particular:

- the methods 1–3 extract Boolean rules, the method 4 fuzzy rules;
- among the Boolean methods, the method 1 extracts classification rules, the method 2 predicate rules with an associational quantifier, and the method 3 predicate rules with an implicational quantifier;
- the methods 2 and 3 are direct methods without an intermediate representation, the method 1 is a classification tree method, and the method 4 is ANN-based:

(1) The method *CART* [38] recursively partitions data with axis-orthogonal hyperplanes, where the choice between different partitions relies on some impurity index, based on estimates $\hat{p}(c|v)$ of the conditional probability that an object in

the vertex v of the partition tree belongs to the class c. For testing, the implementation of CART in MATLAB has been used, with the impurity index being either the Gini index $\sum_{c \neq c'} \hat{p}(c|v)\hat{p}(c'|v)$, or the deviance $-\sum \hat{p}(c|v) \ln \hat{p}(c|v)$.

(2) The *Fisher quantifier* $\sim^F_\alpha$, $\alpha \in (0,1)$ has its truth function $\mathrm{Tf}_{\sim^F_\alpha}$ defined in such a way that the rule $S_r \sim^F_\alpha S'_r$ is valid exactly for those data for which statistical testing of the null hypothesis of independence of S_r and S'_r against the alternative of their positive dependence with the one-tailed Fisher exact test leads to rejecting the null hypothesis on the significance level α [19]. Hence,

$$\mathrm{Tf}_{\sim^F_\alpha}(a,b,c,d) = \begin{cases} 1 & \text{iff } ad > bc \;\&\; \sum_{i=a}^{a+\min(b,c)} \dfrac{\binom{a+c}{i}\binom{b+d}{a+b-i}}{\binom{a+b+c+d}{a+b}} \leqslant \alpha, \\ 0 & \text{else.} \end{cases} \tag{17}$$

For testing, the implementation in the LISP-Miner system [45] was used.

(3) The quantifier *founded implication* $\rightarrow_{s,\theta}$, $s, \theta \in (0,1]$, has its truth function $\mathrm{Tf}_{\rightarrow_{s,\theta}}$ defined in such a way that the rule $A_r \rightarrow_{s,\theta} C_r$ is valid exactly for those data for which the conditional probability $p(C_r|A_r)$ of the validity of C_r conditioned on A_r, estimated with the unbiased estimate $\frac{a}{a+b}$, is at least θ, whereas A_r and C_r are simultaneously valid in at least the proportion s of the data [19]. Hence,

$$\mathrm{Tf}_{\rightarrow_{s,\theta}} = \begin{cases} 1 & \text{iff } \frac{a}{a+b} \geqslant \theta \;\&\; \frac{a}{a+b+c+d} \geqslant s, \\ 0 & \text{else.} \end{cases} \tag{18}$$

As was pointed out in [46], rules with this quantifier are actually association rules with support s and confidence θ. Also in this case, the implementation in LISP-Miner was used for testing.

(4) An ANN-based method for the extraction of rules of any fuzzy propositional logic that was proposed in [47]. It extracts always a single rule $S_r \equiv S'_r$ with atomic S'_r and S_r in disjunctive normal form (DNF), each atom of which contains a single object variable modelled with a finitely-parametrized fuzzy set (e.g., Gaussian, triangular, sigmoid). The architecture of the ANN reflects the construction of the S_r. An example output of such ANN is depicted in Fig. 1. For testing, the method has been implemented in MATLAB [47].

4. Existing measures for classification rulesets

Since sets of classification rules (even rules with fuzzy antecedents) are the output of classification systems, methods assessing the quality of those systems can be employed also as quality measures for classification rulesets. A survey of such measures has been given in the monograph [1]. All measures have been divided there into four groups: inaccuracy, imprecision, inseparability and resemblance. Space limitation allows to recall here only the main representatives of the two more important groups:

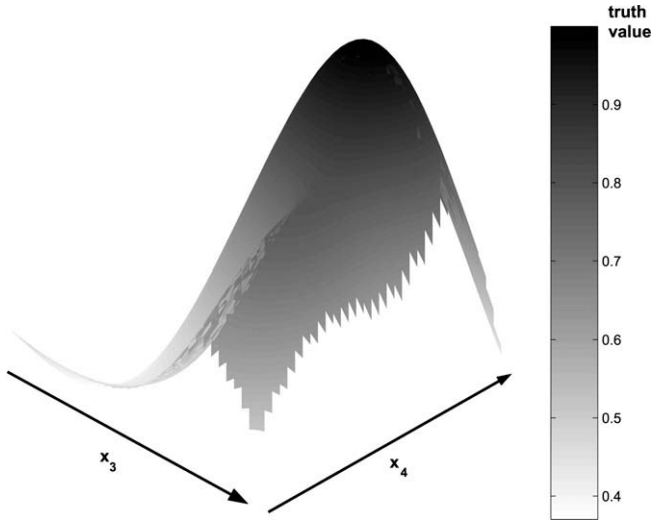

Fig. 1. A 2-dimensional cut for the dimensions x_3 and x_4 of the graph of a mapping computed by a neural network with 12 input neurons, 5 hidden neurons and 1 output neuron, each input of which corresponds to a variable modelled with a Gaussian fuzzy set, whereas the output returns the truth grade of the equivalent formulas.

Inaccuracy measures the discrepancy between the true class of the considered objects and the class predicted by the ruleset. Its most frequently encountered representative is the *quadratic score* (also called Brier score):

$$\mathrm{Inacc} = \frac{1}{|\mathscr{O}|}\sum_{x\in\mathscr{O}}\sum_{C\in\{C_r\}_{r\in\mathscr{R}}}(\delta_C(x)-\hat{\delta}_C(x))^2, \tag{19}$$

where $|\,|$ denotes cardinality, $\mathscr{O}$ is the considered set of objects, $\delta_C(x)\in\{0,1\}$ is the validity of the formula C for a variable x with values in $\mathscr{O}$, and $\hat{\delta}_C(x)$ is the agreement between C and the class predicted for x by $\mathscr{R}$. Hence, $\hat{\delta}_C(x) = \max_{C_r=C}\|A_r\|_x$.

Imprecision measures the discrepancy between the probability distribution of the classes, conditioned on the values of attributes occurring in antecedents, and the class predicted by the ruleset. Its most common representative is

$$\mathrm{Impr} = \frac{1}{|\mathscr{O}|}\sum_{x\in\mathscr{O}}\sum_{C\in\{C_r\}_{r\in\mathscr{R}}}(\delta_C(x)-\hat{\delta}_C(x))(1-\hat{\delta}_C(x))^2. \tag{20}$$

The rulesets that a particular method extracts from given data can substantially depend on values of various parameters of the method, such as tree depth or size for the CART method, significance level for the Fisher quantifier, support and confidence for association rules, or the number of hidden neurons and parameters of the input fuzzy sets for the ANN-based method proposed in [47]. Then also the results of applying quality measures to the ruleset depend on those parameter values. So far, the influence of parameter values has been systematically studied only for dichotomous classification when $\mathscr{R} = \{A\to C, \neg A\to\neg C\}$. In that case, putting $A_r = A, C_r = C$ allows the information about the validity of A and C for $\mathscr{O}$ to be again summarized by means of the 4-fold table (14), which then also depends on the parameter values. The influence of the parameter values on the result of dichotomous classification is usually investigated by means of the measures *sensitivity* $=\frac{a}{a+c}$ and *specificity* $=\frac{d}{b+d}$ [1]. Connecting points (1-specificity,sensitivity) $=(\frac{b}{b+d},\frac{a}{a+c})$ for the considered parameter values forms a curve with graph in the unit square, called *receiver operating characteristic* (ROC), due to the area where such curves have first been in routine use. In machine learning, a modified version of those curves has been proposed, in which the points connected for considered parameter values are (b,a) [2]. The graph of such a curve then lies in the rectangle with vertices $(0,0)$ and $(b+d,a+c)$, and is called *coverage graph*.

The graphs of ROC curves and coverage graphs can provide information about the influence of parameter values not only on the sensitivity and specificity, but also on other measures. It is sufficient to complement the graph with isolines of the measure and to investigate their intersections with the original curve [2].

5. Three different generalizations

In the particular case of classification rulesets with Boolean antecedents, some algebra allows to substantially simplify (19) and (20):

$$\mathrm{Inacc} = \frac{2|\mathscr{O}^-|}{|\mathscr{O}|} = 1-\frac{|\mathscr{O}^+|-|\mathscr{O}^-|}{|\mathscr{O}|},\quad \mathrm{Impr} = \frac{|\mathscr{O}^-|}{|\mathscr{O}|} = 1-\frac{|\mathscr{O}^+|}{|\mathscr{O}|}, \tag{21}$$

where

$$\mathscr{O}^+ = \{x\in\mathscr{O} : \mathscr{R} \text{ is valid for } x\},\quad \mathscr{O}^- = \{x\in\mathscr{O} : \mathscr{R} \text{ is not valid for } x\}. \tag{22}$$

This not only shows that, in the case of Boolean antecedents, $\mathrm{Impr} = \frac{1}{2}\mathrm{Inacc}$, thus the quadratic score is sufficient to describe also the imprecision, but also suggests an approach how to extend those measures to general rulesets: to use (21) and (22) as the definition of measures (19) and (20) in their case. More generally, any measure of quality of classification rulesets with Boolean antecedents (e.g., any measure surveyed in [1]) that can be reformulated by means of $\mathscr{O}^+$ and/or $\mathscr{O}^-$, can be extended in such a way that the reformulation is used as the definition of that measure for general rulesets.

For sets of asymmetric rules, also the notion of covering an object by a rule, which was recalled in Section 3, can be generalized. Notice, however, that for fuzzy antecedents, the validity of A_r, $r\in\mathscr{R}$ is a fuzzy set on $\mathscr{O}$. Consequently, the set $\mathscr{O}_{\mathscr{R}}$ of objects covered by $\mathscr{R}$ is a fuzzy set on $\mathscr{O}$ with the membership function

$$\mu_{\mathscr{R}}(x) = \|(\exists r\in\mathscr{R})A_r\|_x = \max_{r\in\mathscr{R}}\|A_r\|_x. \tag{23}$$

Observe that according to (23), $\mathscr{O}_{\mathscr{R}} = \mathscr{O}$ for classification rulesets with Boolean antecedents. Therefore, various generalizations of classification measures to general rulesets of asymmetric rules are possible: wherever $\mathscr{O}$ occurs in the definition of a measure for classification rulesets, either $\mathscr{O}$ or $\mathscr{O}_{\mathscr{R}}$ can occur in a generalization of that definition, provided $\mathscr{O}_{\mathscr{R}}\neq\emptyset$. For example, both measures

$$\mathrm{Impr}_1 = 1-\frac{|\mathscr{O}^+|}{|\mathscr{O}|}\quad\text{and}\quad \mathrm{Impr}_2 = 1-\frac{|\mathscr{O}^+|}{|\mathscr{O}_{\mathscr{R}}|}, \tag{24}$$

where in general

$$|\mathscr{O}_{\mathscr{R}}| = \sum_{x\in\mathscr{O}}\mu_{\mathscr{R}}(x) \tag{25}$$

are generalizations of (20).

Observe that if $|\mathcal{O}_{\mathscr{R}}| = |\mathcal{O}|$, then $\mathrm{Impr}_2 = \mathrm{Impr}_1$. Moreover, if the stronger condition

$$|\mathcal{O}^+| + |\mathcal{O}^-| = |\mathcal{O}_{\mathscr{R}}| = |\mathcal{O}| \tag{26}$$

holds, then the relationships between the inaccuracy and imprecision measures in (38) and (39) becomes as simple as in (21):

$$\mathrm{Impr}_1 = \mathrm{Impr}_2 = \frac{1}{2}\mathrm{Inacc}. \tag{27}$$

To allow unified treatment of symmetric and asymmetric rules, the concept of covering an object by a rule will be extended also to symmetric rules, in such a way that an object x is covered by $S_r \sim S'_r$ if either S_r or S'_r is valid for x. Hence, a counterpart of (23) for a set $\mathscr{R}$ of symmetric rules is a fuzzy set with the membership function

$$\mu_{\mathscr{R}}(x) = \|(\exists r \in \mathscr{R})(S_r \vee S'_r)\|_x = \max_{r\in\mathscr{R}} \max(\|S_r\|_x, \|S'_r\|_x). \tag{28}$$

According to (22), the proposed way of extending measures of quality from classification rulesets with Boolean antecedents to general rulesets requires to generalize the concept of validity of a general ruleset for an object. However, there are multiple possibilities for such a generalization. Indeed, at least any of the following three approaches is feasible.

5.1. Boolean validity of a ruleset based on simultaneous validity of all covering rules

According to this approach, the validity of a ruleset $\mathscr{R}$ for a covered object x is a Boolean property expressing the simultaneous validity of all rules that cover x. Consequently, the sets $\mathcal{O}^+$ and $\mathcal{O}^-$ defined in (22) are crisp sets

$$\mathcal{O}^+ = \{x \in \mathcal{O} : \mu_{\mathscr{R}}(x) > 0 \,\&(\forall r \in \mathscr{R})\|r \text{ covers } x \& r \text{ is valid for } x\| = \|r \text{ covers } x\|\}, \tag{29}$$

$$\mathcal{O}^- = \{x \in \mathcal{O} : \mu_{\mathscr{R}}(x) > 0 \&(\exists r \in \mathscr{R})\|r \text{ covers } x \& r \text{ is valid for } x\| < \|r \text{ covers } x\|\}, \tag{30}$$

$$\text{where } \|r \text{ covers } x\| = \begin{cases} \|(S_r \vee S'_r)\|_x & \text{for symmetric rules,} \\ \|A_r\|_x & \text{for asymmetric rules,} \end{cases} \tag{31}$$

$$\text{and similarly } \|r \text{ covers } x \,\&\, r \text{ is valid for } x\| = \begin{cases} \|(S_r \vee S'_r)\&r\|_x & \text{for symmetric rules,} \\ \|A_r\&r\|_x & \text{for asymmetric rules.} \end{cases} \tag{32}$$

The following consequences of this approach are worth noticing:

(i) It is immaterial how the truth grade $\|r\|_x$ of a rule r being valid for an object x is evaluated (thus also how $\|\neg r\|_x$ is evaluated).
(ii) If $\mu_{\mathscr{R}}(x) = 0$, then $x \notin \mathcal{O}^+ \cup \mathcal{O}^-$.
(iii) For classification rulesets with Boolean antecedents, the validity of $\mathscr{R}$ for an object x coincides with the definition in Section 3 because in that case, there is exactly one rule that covers x.

5.2. Boolean validity of a ruleset based on the validity of the majority of covering rules

According to this approach, the validity of a ruleset $\mathscr{R}$ for a covered object x is a Boolean property expressing the validity of most of the rules that cover x. Consequently, the sets $\mathcal{O}^+$ and $\mathcal{O}^-$ in (22) are crisp sets

$$\mathcal{O}^+ = \left\{x \in \mathcal{O} : \mu_{\mathscr{R}}(x) > 0 \,\&\, \sum_{r\in\mathscr{R}} \|r \text{ covers } x \,\&\, r \text{ is valid for } x\| > \sum_{r\in\mathscr{R}} \|r \text{ covers } x \,\&\, \neg\, r \text{ is valid for } x\|\right\}, \tag{33}$$

$$\mathcal{O}^- = \left\{x \in \mathcal{O} : \mu_{\mathscr{R}}(x) > 0 \,\&\, \sum_{r\in\mathscr{R}} \|r \text{ covers } x \,\&\, r \text{ is valid for } x\| \leqslant \sum_{r\in\mathscr{R}} \|r \text{ covers } x \,\&\, \neg\, r \text{ is valid for } x\|\right\}, \tag{34}$$

where the truth grade $\|r \text{ covers } \&\, \neg\, r \text{ is valid for } x\|$ is again evaluated according to (32), replacing r with $\neg r$. Observe that also this approach has the above consequences (i)–(iii), the last one again due to the fact that there is exactly one rule covering x.

5.3. Fuzzy validity of a ruleset based on the relative validity of covering rules

In this case, the validity of a ruleset $\mathscr{R}$ for a covered object x is a fuzzy property expressing the ratio of the validity of rules from $\mathscr{R}$ for x to the covering of x with those rules. Consequently, the sets $\mathcal{O}^+$ and $\mathcal{O}^-$ are fuzzy sets on $\mathcal{O}$ with memberships μ_+ and μ_-, respectively, such that if $\mu_{\mathscr{R}}(x) > 0$,

$$\mu_+(x) = \frac{\sum_{r\in\mathscr{R}}\|r \text{ covers } x \,\&\, r \text{ is valid for } x\|}{\sum_{r\in\mathscr{R}}\|r \text{ covers } x\|\|}, \tag{35}$$

$$\mu_-(x) = \frac{\sum_{r\in\mathscr{R}}\|r \text{ covers } x \,\&\, \neg\, r \text{ is valid for } x\|}{\sum_{r\in\mathscr{R}}\|r \text{ covers } x\|\|}, \tag{36}$$

where the involved truth grades are again evaluated according to (31) and (32). Moreover, (35) and (36) will be complemented with the definition $\mu_+(x) = \mu_-(x) = 0$ if $\mu_{\mathscr{R}}(x) = 0$, to get again the validity of (ii) above, whereas (i) and (iii) are consequences also of this approach.

The fact that $\mathscr{O}_{\mathscr{R}}$, $\mathscr{O}^+$ and $\mathscr{O}^-$ are fuzzy sets implies that whenever $|\mathscr{O}_{\mathscr{R}}|$, $|\mathscr{O}^+|$ or $|\mathscr{O}^-|$ occur in the definitions of quality measures for Boolean classification rulesets, fuzzy cardinalities have to be used in their extensions to general rulesets. Consequently,

$$|\mathscr{O}^+| = \sum_{x\in\mathscr{O}}\mu_+(x), \quad |\mathscr{O}^-| = \sum_{x\in\mathscr{O}}\mu_-(x). \tag{37}$$

Hence, the measure (19) now turns to

$$\text{Inacc} = 1 - \frac{\sum_{x\in\mathscr{O}}(\mu_+(x) - \mu_-(x))}{|\mathscr{O}|}, \tag{38}$$

due to (22), whereas (25) turns to

$$\text{Impr}_1 = 1 - \frac{\sum_{x\in\mathscr{O}}\mu_+(x)}{|\mathscr{O}|}, \quad \text{Impr}_2 = 1 - \frac{\sum_{x\in\mathscr{O}}\mu_+(x)}{\sum_{x\in\mathscr{O}}\mu_{\mathscr{R}}(x)}. \tag{39}$$

6. Extensions of ROC curves to more general kinds of rules

Observe that in the case of Boolean classification with $\mathscr{R} = \{A \to C, \neg A \to \neg C\}$, the information about the validity of $\mathscr{R}$ for objects $x \in \mathscr{O}$ can be also viewed as information about the validity of a ruleset $\mathscr{R}' = \{A \to C\}$. However, $\mathscr{R}'$ is not any more a classification ruleset, but only a general one, which can be described only by means of the above introduced sets $\mathscr{O}_{\mathscr{R}}$, $\mathscr{O}^+$, $\mathscr{O}^-$. In particular, $|\mathscr{O}^+| = a$ and $|\mathscr{O}^-| = b$, which suggests the possibility to generalize coverage graphs introduced in Section 4 to general rulesets by means of a curve connecting points $(|\mathscr{O}^-|, |\mathscr{O}^+|)$ for each of the considered parameter values. For a generalization of ROC curves to general rulesets, those points have to be scaled to the unit square. Since the resulting curve will be used to investigate the dependence on parameter values, the scaling factor itself must be independent of those values. The only available factor fulfilling this condition is the number of objects, $|\mathscr{O}|$ (the other available factors, $|\mathscr{O}_{\mathscr{R}}|$, $|\mathscr{O}^+|$ and $|\mathscr{O}^-|$ depend on the truth evaluations $\|S_r\|$ and $\|S'_r\|$, or $\|A_r\|$ and $\|C_r\|$, which in turn depend on the parameter values). Consequently, the proposed generalization of ROC curves will connect points $\left(\frac{|\mathscr{O}^-|}{|\mathscr{O}|}, \frac{|\mathscr{O}^+|}{|\mathscr{O}|}\right)$.

For the applicability of such a generalization of ROC curves, the following proposition can be quite useful:

Proposition 1. *Let the covering of individual objects with individual rules be a Boolean property (i.e., the set of rules covering a particular object x be a crisp subset of $\mathscr{R}$). Then irrespectively of which of the above approaches to ruleset validity is adopted, there always exists a constant $c \in (0,1]$ and a strictly increasing bijective mapping $g : [0,c] \to [0,1]$ such that*

$$|\mathscr{O}^+| + |\mathscr{O}^-| \leqslant \max\left(\max_{x\in[0,c]}(x + g^{-1}(1 - g(x))), 1\right)|\mathscr{O}|. \tag{40}$$

Moreover, in the particular cases of Boolean logic and of all three fundamental fuzzy logics (Gödel, Łukasiewicz, product), (40) holds with $c = 1$ and g equal to identity:

$$|\mathscr{O}^+| + |\mathscr{O}^-| \leqslant |\mathscr{O}|. \tag{41}$$

Thus in those cases, the points $\left(\frac{|\mathscr{O}^-|}{|\mathscr{O}|}, \frac{|\mathscr{O}^+|}{|\mathscr{O}|}\right)$ forming the generalization of ROC curves, lie below the diagonal $([0,1],[1,0])$.

Proof. Denote $\mathscr{R}_x = \{r \in \mathscr{R} : r \text{ covers } x\}$. According to the assumption of the proposition, $\mathscr{R}_x$ is crisp. Putting a crisp set into (qmp)–(qmn) simplifies them to

$$\mu_+(x) = \frac{\sum_{r\in\mathscr{R}_x}\|r\|}{|R_x|}, \quad \mu_-(x) = \frac{\sum_{r\in\mathscr{R}_x}\|\neg r\|}{|R_x|}. \tag{42}$$

Combining (42) with (37) and with Lemma 2 applied to the choice $\varphi = r \in \mathscr{R}$ already yields the inequality (40). Similarly, applying Lemma 1(e) to the choice $\varphi = r \in \mathscr{R}$ gives

$$\|r\|_t + \|\neg r\|_t \leqslant 1, \tag{43}$$

which in combination with (37) and (42) entails the inequality (41). □

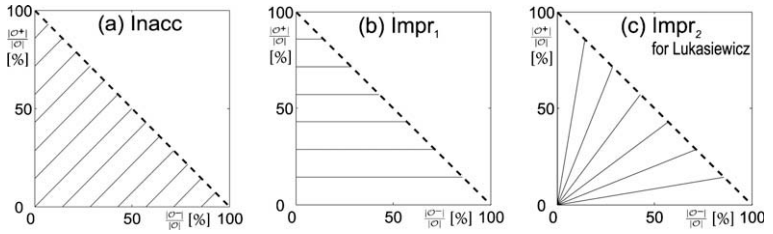

Fig. 2. Isolines of the three measures introduced in (38) and (39), drawn with respect to the coordinates $\left(\frac{|\mathscr{O}^-|}{|\mathscr{O}|}, \frac{|\mathscr{O}^+|}{|\mathscr{O}|}\right)$ of points forming the proposed generalization of ROC curves.

The proposition can be useful through delimiting the area where the generalized ROC curve can lie, in particular if isolines of the considered measure are constructed in that area. This is illustrated in Fig. 2, together with isolines of the three example measures introduced in (38) and (39). Observe that the isolines of Impr_2 depend on the relationship between the three cardinalities $|\mathscr{O}^+| = \sum_{x\in\mathscr{O}} \mu_+(x)$, $|\mathscr{O}^-| = \sum_{x\in\mathscr{O}} \mu_-(x)$ and $|\mathscr{O}_{\mathscr{R}}| = \sum_{x\in\mathscr{O}} \mu_{\mathscr{R}}(x)$. The isolines depicted in Fig. 2(c) correspond to the relationship $|\mathscr{O}_{\mathscr{R}}| = |\mathscr{O}^+| + |\mathscr{O}^-|$, which is true in Łukasiewicz logic (and in particular also in Boolean logic). In the delimited area, the intersections of the isolines with the curve can be subsequently searched, like in the case of traditional ROC curves. In addition, that area allows to obtain an upper bound to the *area under the curve* (AUC), which can serve as a particular quality measure [1,18].

7. Illustration using Fisher iris data

The three approaches proposed in Section 5 were so far tested for six rules extraction methods, including the four methods recalled in Section 3. Testing was performed on three benchmark data sets, as well as on data from one real-world knowledge discovery task [48]. The results of testing will now be illustrated on the best known benchmark set, the iris data, originally used in 1930s by Fisher [49].

As to the methods recalled in Section 3, the method CART was used with trees of 2–6 leaves, each combined with Gini index and deviance, the Fisher quantifier with 5 significance levels, the founded implication with combinations of 10 values of s and 7 values of θ, and the ANN-based method from [47] with combinations of 2–4 hidden neurons and 3 particular fuzzy sets modelling input variables, each of them interpreted in Łukasiewicz and in product-Łukasiewicz logic. All three approaches were tested for the considered methods extracting Boolean rules, i.e., the method CART, the founded implication and the Fisher quantifier. In addition, the fuzzy approach described in Section 5.3 was tested also for the ANN-based method extracting fuzzy logic rules.

Whereas each rule extracted with the ANN-based method defines a specific atomic fuzzy concept described with the equivalent DNF, the three Boolean methods were used to extract traditional rules for the iris data, concerning relationships between the values of their descriptive attributes (length and width of petals and sepals) and the kind of iris, i.e., such rules as $1\ \mathrm{cm} \leqslant \mathtt{petal\ length} \leqslant 3\ \mathrm{cm} \rightarrow \mathtt{Setosa}$, or $1\ \mathrm{cm} \leqslant \mathtt{petal\ length} \leqslant 3\ \mathrm{cm}\ \&\ 1\ \mathrm{mm} \leqslant \mathtt{petal\ width} \leqslant 6\ \mathrm{mm} \sim \mathtt{Setosa}$. The fact that in each such rule, C_r, S_r or S_r is an assignment to values of the classification attribute kind of iris could superficially lead to the impression that all three methods actually extracted classification rules from the iris data. Needless to say, that impression is false: recall from Section 3 that among the considered methods, only CART extracts classification rules, whereas the rules extracted by the other three methods are more general. To decrease the probability that such an impression can arise, also rules concerning relationships between the values of different descriptive attributes were extracted by means of the Fisher quantifier. Examples of such extracted rules were:

$$4.8\ \mathrm{cm} \leqslant \mathtt{petal\ length} \leqslant 6.7\ \mathrm{cm}\ \&\ 1\ \mathrm{mm} \leqslant \mathtt{petal\ width} \leqslant 25\ \mathrm{mm} \sim^{\mathrm{F}}_{0.1\%} 6.3\ \mathrm{cm} \leqslant \mathtt{sepal\ length} \leqslant 7.9\ \mathrm{cm},$$
$$4.8\ \mathrm{cm} \leqslant \mathtt{petal\ length} \leqslant 6.7\ \mathrm{cm} \sim^{\mathrm{F}}_{1\%} 6.3\ \mathrm{cm} \leqslant \mathtt{sepal\ length} \leqslant 7.9\ \mathrm{cm},$$
$$1\ \mathrm{mm} \leqslant \mathtt{petal\ width} \leqslant 3\ \mathrm{mm} \sim^{\mathrm{F}}_{5\%} 31\ \mathrm{mm} \leqslant \mathtt{sepal\ width} \leqslant 44\ \mathrm{mm}.$$

For the split of the data into training and test set, a 10-fold cross validation was employed. Consequently, always $10 * (2 * 5 + 2 * 5 + 10 * 7) = 900$ rulesets were extracted with the considered methods from the iris data using the Boolean approaches described in Sections 5.1 and 5.2, whereas $900 + 10 * 2 * 3 * 3 = 1080$ rulesets were extracted using the fuzzy approaches described in Section 5.3.

Results obtained when applying the measures introduced in (38) and (39) to the extracted rulesets are given in Tables 1–3. When using cross validation, the reported results are average values over all folds of evaluating the ruleset quality with that part of data that served as test data in the respective fold (whereas serving as training data in the remaining folds). For the founded implication only 10 example combinations of the values of s and θ have been included, from among the 70 combinations with which the method was tested. For that method, the obtained results of all three measures, Inacc, Impr_1, and Impr_2 are given. In all other case, the values of Impr_1 are sufficient. The number of rules in a CART ruleset equals the number of leaves of the tree, and with the ANN-based method, always one DNF rule is extracted. On the other hand, the

Table 1
Average results obtained for the rulesets extracted from the iris data with Boolean methods from Section 3 using the approach described in Section 5.1.

CART									
Leaves				2	3	4	5	6	
Gini index			$Impr_1$	0.40	0.08	0.04	0.02	0.16	
Deviance			$Impr_1$	0.40	0.08	0.04	0.02	0.16	
Fisher quantifier									
Significance level α (%)				0.1	0.5	1	5	10	
Description vs. kind of iris			$Impr_1$	1.0	1.0	1.0	1.0	1.0	
Only descriptive attributes			$Impr_1$	1.0	1.0	1.0	1.0	1.0	
Founded implication									
s	θ	Inacc	$Impr_1$	$Impr_2$	s	θ	Inacc	$Impr_1$	$Impr_2$
0.03	0.8	0.38	0.19	0.19	0.07	0.8	0.28	0.28	0.14
0.03	0.9	0.21	0.11	0.10	0.07	0.85	0.21	0.11	0.10
0.05	0.8	0.36	0.18	0.18	0.09	0.6	1.0	0.5	0.5
0.05	0.85	0.25	0.13	0.12	0.09	0.7	0.43	0.22	0.21
0.05	0.9	0.21	0.11	0.10	0.09	0.8	0.17	0.10	0.07

Table 2
Average results obtained for the rulesets extracted from the iris data with Boolean methods from Section 3 using the approach described in Section 5.2.

CART									
Leaves				2	3	4	5	6	
Gini index			$Impr_1$	0.40	0.08	0.04	0.02	0.16	
Deviance			$Impr_1$	0.40	0.08	0.04	0.02	0.16	
Fisher quantifier									
Significance level α (%)				0.1	0.5	1	5	10	
Description vs. kind of iris			$Impr_1$	0.0	0.0	0.0	0.0	0.0	
Only descriptive attributes			$Impr_1$	0.0	0.0	0.0	0.0	0.0	
Founded implication									
s	θ	Inacc	$Impr_1$	$Impr_2$	s	θ	Inacc	$Impr_1$	$Impr_2$
0.03	0.8	0.06	0.03	0.03	0.07	0.8	0.10	0.05	0.05
0.03	0.9	0.07	0.04	0.03	0.07	0.85	0.07	0.04	0.03
0.05	0.8	0.06	0.03	0.03	0.09	0.6	0.14	0.07	0.07
0.05	0.85	0.05	0.03	0.02	0.09	0.7	0.09	0.05	0.04
0.05	0.9	0.07	0.04	0.03	0.09	0.8	0.07	0.05	0.02

Table 3
Average results obtained for the rulesets extracted from the iris data with methods from Section 3 using the approach described in Section 5.3.

CART									
Leaves				2	3	4	5	6	
Gini index			$Impr_1$	0.40	0.08	0.04	0.02	0.16	
Deviance			$Impr_1$	0.40	0.08	0.04	0.02	0.16	
Fisher quantifier									
Significance level α (%)				0.1	0.5	1	5	10	
Description vs. kind of iris			$Impr_1$	0.24	0.25	0.26	0.27	0.27	
Only descriptive attributes			$Impr_1$	0.21	0.21	0.22	0.22	0.22	
Founded implication									
s	θ	Inacc	$Impr_1$	$Impr_2$	s	θ	Inacc	$Impr_1$	$Impr_2$
0.03	0.8	0.14	0.07	0.07	0.07	0.8	0.12	0.06	0.06
0.03	0.9	0.11	0.06	0.05	0.07	0.85	0.11	0.06	0.05
0.05	0.8	0.12	0.06	0.06	0.09	0.6	0.32	0.16	0.16
0.05	0.85	0.11	0.06	0.05	0.09	0.7	0.21	0.11	0.10
0.05	0.9	0.09	0.05	0.04	0.09	0.8	0.11	0.07	0.04

ANN-based method [47]

Logic		Łukasiewicz			Product-Łukasiewicz		
Number of hidden neurons		2	3	4	2	3	4
Gaussian input	$Impr_1$	0.04	0.05	0.04	0.69	0.69	0.69
Triangular input		0.20	0.20	0.18	0.11	0.18	0.18
Sigmoid input		0.30	0.25	0.25	0.69	0.69	0.68

Table 4
Average number of rules in the rulesets extracted from the iris data with the Fisher quantifier and the founded implication.

Fisher quantifier					
Rules at significance level α (%)	0.1	0.5	1	5	10
Description vs. kind of iris	77.4	104.6	111.8	151	158
Only descriptive attributes	159.2	228.6	277.6	427.6	520.8

s	θ	Rules	s	θ	Rules
Founded implication					
0.03	0.8	100	0.07	0.8	45.7
0.03	0.9	83.5	0.07	0.85	43.1
0.05	0.8	71.7	0.09	0.6	35.2
0.05	0.85	66.5	0.09	0.7	31.6
0.05	0.9	59.8	0.09	0.8	29.3

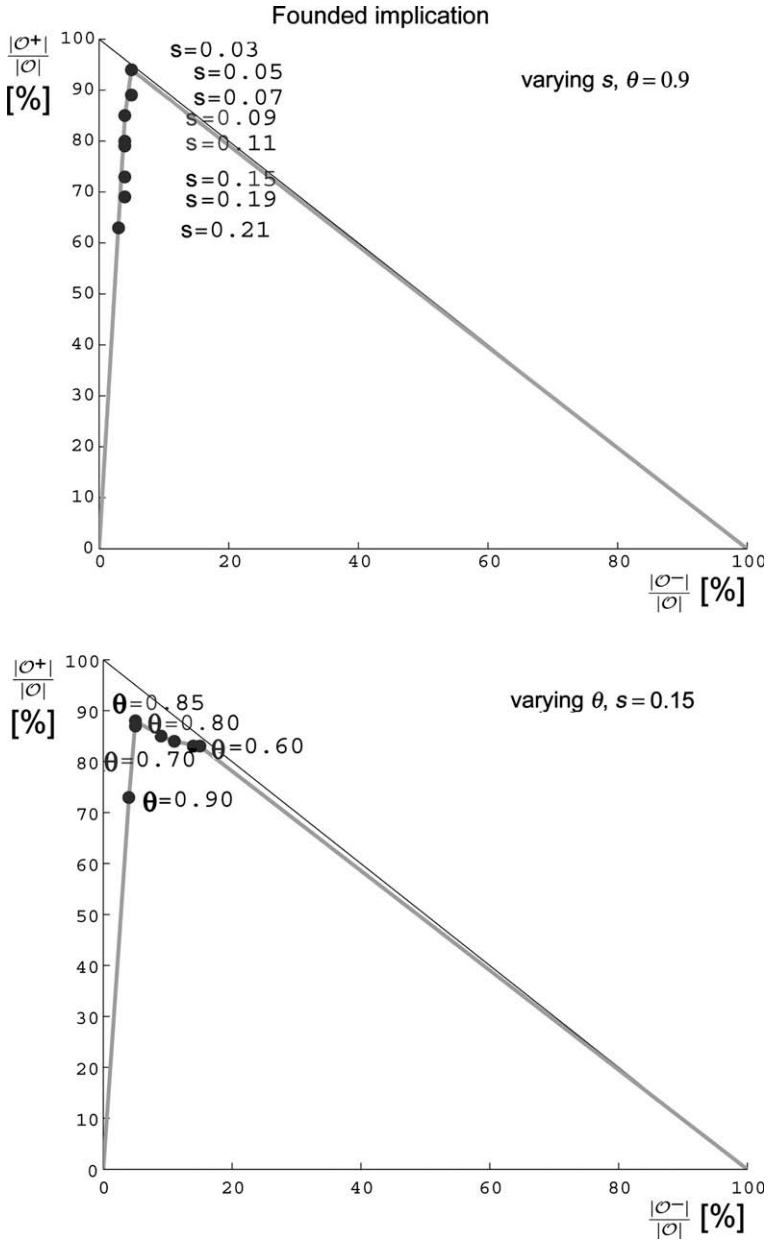

Fig. 3. Generalized ROC curves for rulesets extracted from the iris data by means of the founded implication with $\theta = 0.9$ and a varying s (top), and with $s = 0.15$ and a varying θ (bottom).

size of rulesets extracted using the Fisher quantifier or the founded implication can vary considerably. Therefore, the number of extracted rules, again averaged over the 10-fold cross validation, has been reported for those two methods in Table 4.

Finally, Fig. 3 shows two examples of the proposed generalization of ROC curves for rules extracted from the iris data by means of the founded implication, both with a fixed θ and a varying s, and with a fixed s and a varying θ. The coordinates $\left(\frac{|\mathcal{O}^-|}{|\mathcal{O}|}, \frac{|\mathcal{O}^+|}{|\mathcal{O}|}\right)$ in this figure were computed using the approach described in Section 5.3.

Apart from confirming the expectation that the most precise assignment of the classification attribute kind of iris is achieved with a specific classification method, i.e., the method CART, the presented results support the following observations:

(i) The approach based on simultaneous validity of all covering rules, described in Section 5.1, is sometimes too strict, in the sense that the obtained values of inaccuracy and imprecision are unrealistically high. On the other hand, the approach based on the majority of covering rules, described in Section 5.2 is sometimes too weak, in the sense that the obtained values of inaccuracy and imprecision are unrealistically low. Finally, the values obtained using the fuzzy approach described in Section 5.3 not only never exceed those obtained using the approach based on simultaneous validity of all covering rules (which is a consequence of definitions in Section 5), but also never fell below the values obtained using the majority approach (which is not their necessary consequence). Even more importantly, the obtained values of inaccuracy and imprecision do not seem unrealistically high or low.

(ii) The choice of values of method parameters had a much greater impact on the ruleset quality than the choice of the method itself. Whereas even methods relying on quite different theoretical principles yielded rulesets of comparable quality, inappropriate values of parameters turned the method from a useful one to a quite useless one.

8. Conclusions

The paper has dealt with quality measures of rules extracted from data, though not in the usual context of individual rules, but in the context of whole rulesets. Three kinds of extensions of measures already in use for classification rulesets have been proposed and example results of extensive tests on rulesets extracted with four important data mining methods have been presented. For all three proposed extensions, the validity of a ruleset for an object coincides with the traditional definition if a classification ruleset with Boolean antecedents is considered. In addition, the concept of ROC curves has been generalized, to enable investigating the dependence of general rulesets on parameter values of the extraction method.

The extent of the paper did not allow more than only to sketch the basic ideas of the proposed approaches and to include results for the best known benchmark data set, Fisher iris data. Nevertheless, even these results alone indicate that the approach is feasible and can contribute to the ultimate objective of quality measures: to allow comparing the knowledge extracted with different data mining methods and investigating how the extracted knowledge depends on the values of their parameters.

References

[1] D. Hand, Construction and Assessment of Classification Rules, John Wiley and Sons, New York, 1997.
[2] J. Fürnkranz, P. Flach, ROC 'n' rule learning – towards a better understanding of covering algorithms, Machine Learning 58 (2005) 39–77.
[3] K. Kaufman, R. Michalski, An adjustable description quality measure for pattern discovery using the AQ methodology, Journal of Intelligent Information Systems 14 (2000) 199–216.
[4] S. Greco, Z. Pawlak, R. Slowinski, Can Bayesian confirmation measures be useful for rough set decision rules, Engineering Applications of Artificial Intelligence 20 (2004) 345–361.
[5] K. McGarry, A survey of interestingness measures for knowledge discovery, Knowledge Engineering Review 20 (2005) 39–61.
[6] D. Dubois, Hüllermeier, H. Prade, A systematic approach to the assessment of fuzzy association rules, Data Mining and Knowledge Discovery 13 (2006) 167–192.
[7] I. Brzezinska, S. Greco, R. Slowinski, Mining pareto-optimal rules with respect to support and confirmation or support and anti-support, Engineering Applications of Artificial Intelligence 20 (2007) 587–600.
[8] L. Geng, H. Hamilton, Choosing the right lens: finding what is interesting in data mining, in: F. Guillet, H. Hamilton (Eds.), Quality Measures in Data Mining, Springer Verlag, Berlin, 2007, pp. 3–24.
[9] S. Lallich, O. Teytaud, E. Prudhomme, Association rule interestingness: measure and statistical validation, in: F. Guillet, H. Hamilton (Eds.), Quality Measures in Data Mining, Springer Verlag, Berlin, 2007, pp. 251–275.
[10] S. Bay, M. Pazzani, Detecting group differences. mining contrast sets, Data Mining and Knowledge Discovery 5 (2001) 213–246.
[11] R. Hilderman, T. Peckham, Statistical methodologies for mining potentially interesting contrast sets, in: F. Guillet, H. Hamilton (Eds.), Quality Measures in Data Mining, Springer Verlag, Berlin, 2007, pp. 153–177.
[12] L. Lerman, J. Azè, Une mesure probabiliste contextuelle discriminante de qualite des règles d'association, in: EGC 2003: Extraction et Gestion des Connaissances, Hermes Science Publications, Lavoisier, 2003, pp. 247–263.
[13] P. Lenca, B. Vaiilant, P. Meyer, S. Lalich, Association rule interestingness meaures: experimental and theoretical studies, in: F. Guillet, H. Hamilton (Eds.), Quality Measures in Data Mining, Springer Verlag, Berlin, 2007, pp. 51–76.
[14] M. Holeňa, Measures of ruleset quality capable to represent uncertain validity, in: K. Mellouli (Ed.), ECSQARU 2007: Ninth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer Verlag, Berlin, 2007, pp. 430–442.
[15] V. Novák, I. Perfilieva, J. Močkoř, Mathematical Principles of Fuzzy Logic, Kluwer Academic Publishers, Dordrecht, 1999.
[16] P. Hájek, Metamathematics of Fuzzy Logic, Kluwer Academic Publishers, Dordrecht, 1998.
[17] E. Klement, R. Mesiar, E. Pap, Triangular Norms, Kluwer Academic Publishers, Dordrecht, 2000.
[18] L. Peterson, M. Coleman, Machine learning based receiver operating characteristic (ROC) curves for crisp and fuzzy classification of DNA microarrays in cancer research, International Journal of Approximate Reasoning 47 (2008) 17–36.

[19] P. Hájek, T. Havránek, Mechanizing Hypothesis Formation, Springer Verlag, Berlin, 1978.
[20] M. Zaki, S. Parathasarathy, M. Ogihara, W. Li, New parallel algorithms for fast discovery of association rules, Data Mining and Knowledge Discovery 1 (1997) 343–373.
[21] F. Korn, A. Labrinidis, Y. Kotidis, C. Faloutsos, Quantifiable data mining using ration rules, VLDB Journal 8 (2000) 254–266.
[22] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. Verkamo, Fast discovery of association rules, in: Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, 1996, pp. 307–328.
[23] R. Michalski, Knowledge acquisition through conceptual clustering: a theoretical framework and algorithm for partitioning data into conjunctive concepts, International Journal of Policy Analysis and Information Systems 4 (1980) 219–243.
[24] R. Michalski, K. Kaufman, Learning patterns in noisy data, in: Machine Learning and Its Applications, Springer Verlag, New York, 2001, pp. 22–38.
[25] P. Clark, R. Boswell, Rule induction with CN2: some recent improvements, in: Machine Learning – EWSL-91, Springer Verlag, New York, 1991, pp. 151–163.
[26] Z. Pawlak, Rough Sets, Kluwer Academic Publishers, Dordrecht, 1991.
[27] Y. Leung, M. Fischer, W. Wu, J. Mi, A rough set approach for the discovery of classification rules in interval-valued information systems, International Journal of Approximate Reasoning 47 (2008) 233–246.
[28] L. De Raedt, Interactive Theory Revision: An Inductive Logic Programming Approach, Academic Press, London, 1992.
[29] S. Muggleton, Inductive Logic Programming, Academic Press, London, 1992.
[30] A. Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer Verlag, Berlin, 2002.
[31] J. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Transactions on Systems, Man, and Cybernetics 23 (1993) 665–685.
[32] J. Jang, C. Sun, Neuro-fuzzy modeling and control, The Proceedings of the IEEE 83 (1995) 378–406.
[33] D. Nauck, R. Kruse, NEFCLASS-X: a neuro-fuzzy tool to build readable fuzzy classifiers, BT Technology Journal 3 (1998) 180–192.
[34] D. Nauck, Fuzzy data analysis with NEFCLASS, International Journal of Approximate Reasoning 32 (2002) 103–130.
[35] M. Holeňa, Fuzzy hypotheses for GUHA implications, Fuzzy Sets and Systems 98 (1998) 101–125.
[36] M. Holeňa, Fuzzy hypotheses testing in the framework of fuzzy logic, Fuzzy Sets and Systems 145 (2004) 229–252.
[37] V. Novák, I. Perfilieva, A. Dvořák, C. Chen, Q. Wei, P. Yan, Mining pure linguistic associations from numerical data, International Journal of Approximate Reasoning 48 (2008) 4–22.
[38] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth, Belmont, 1984.
[39] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Francisco, 1992.
[40] W. Duch, R. Adamczak, K. Grabczewski, A new methodology of extraction, optimization and application of crisp and fuzzy logical rules, IEEE Transactions on Neural Networks 11 (2000) 277–306.
[41] H. Tsukimoto, Extracting rules from trained neural networks, IEEE Transactions on Neural Networks 11 (2000) 333–389.
[42] M. Holeňa, Piecewise-linear neural networks and their relationship to rule extraction from data, Neural Computation 18 (2006) 2813–2853.
[43] A. Tickle, R. Andrews, M. Golea, J. Diederich, The truth will come to light: directions and challenges in extracting rules from trained artificial neural networks, IEEE Transactions on Neural Networks 9 (1998) 1057–1068.
[44] S. Mitra, Y. Hayashi, Neuro-fuzzy rule generation: survey in soft computing framework, IEEE Transactions on Neural Networks 11 (2000) 748–768.
[45] M. Šimùnek, Academic KDD project LISP-Miner, in: A. Abraham, K. Franke, K. Koppen (Eds.), Advances in Soft Computing – Systems Design and Applications, Springer Verlag, Heidelberg, 2003, pp. 263–272.
[46] P. Hájek, M. Holeňa, Formal logics of discovery and hypothesis formation by machine, Theoretical Computer Science 292 (2003) 345–357.
[47] M. Holeňa, Extraction of fuzzy logic rules from data by means of artificial neural networks, Kybernetika 41 (2005) 297–314.
[48] M. Holeňa, Neural networks for extraction of fuzzy logic rules with application to EEG data, in: B. Ribeiro, R. Albrecht, A. Dobnikar (Eds.), Adaptive and Natural Computing Algorithms, Springer Verlag, Wein, 2005, pp. 369–372.
[49] R. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 179–188.